PREDICTION OF MEDIAN VALUE OF HOUSE IN BOSTON SUBURBS

A PROJECT REPORT

Presented to the Department of Mathematics and Statistics

California State University, Long Beach

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Mathematics

Option in Statistics

Faculty Reviewer:

Kagba Suaray, Ph.D.

Junyan Deng

NOV 2015

Table of Contents

1 Introduction

Median value is a good measurement to evaluate the house value in a specific area. This median value is related to some elements which are predictors. Understanding what important elements can predict the median value, city officers and the residents can improve these elements to better their house value. The aim of this paper is to determine what important elements can impact the median value, and predict a median value for a specific data set.

2 Data

The data were collected by Harrison, D. and Rubinfeld, D.L in 1993, and is taken from the StatLib library which is maintained at Carnegie Mellon University. The data consists of 506 observations and 14 attributes. Figure 1 displays the table of the attributes and a brief description of each variable. The independent variable used in this analysis will be MADV(median value) that is measured in $1000. I split the data into a model building set and a validation set using PROC SURVEYSELECT using 70% of the data into the model building set and 30% into the validation set.

Table 1-the data

1.CRIM	continuous	Per capita crime rate by town
2 ZN	continuous	Proportion of residential land zoned for lots over 25,000 sq.ft
3 INDUS	continuous	proportion of non-retail business acres per town
4 CHAS	indicator	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5 NOX	continuous	nitric oxides concentration (parts per 10 million)
6 RM	integer	average number of rooms per dwelling
7 AGE	continuous	proportion of owner-occupied units built prior to 1940
8 DIS	allocated codes	weighted distances to five Boston employment centers
9 RAD	allocated	index of accessibility to radial highways
10 TAX	continuous	full-value property-tax rate per $10,000
11 PTRATIO	continuous	pupil-teacher ratio by town
12 B	continuous	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13 LSTAT	continuous	Proportion of lower status of the population
14 1MEDV	continuous	Median value of owner-occupied homes in $1000's

3 Methodology

Multiple linear regression is used to determine which subset of predictor variables are significant. I will make a correlation transformation to get an overview importance of the predictor variables. To reduce multicollinearity of higher order variables, I will center the predictor variables by subtracting the mean for each variable. Several different methods, such as, different criterions and stepwise method, will be used to find the optimal model. Weighted least square is used to cancel the influence of outliers. Once reach a “best” model, then ensure the model is free of multicollinearity and conduct diagnostic check, the residual analysis, for any violations of the underlying assumptions. Finally, apply the model to the validation set to check for any bias that may be in the model. If the model has little bias, then interpret the parameters of the final model and their confident intervals, and draw a conclusion from data analysis.

4 Variables correlation

Table 2 shows the correlation between variables. The left side column show the relation between dependent variable and dependent variable. Unfortunately all the variable are highly related to the dependent variable. This implies multicollinearity. So centering the dependent variables is necessary to minimize the effect of multicollinearity.

When looking at the data, we can find that x13 and x6 are the most correlated to the dependent. The signs of the association helps us know the effect of independent variable on the dependent variable. These help us to build the model. X1, x3, x5, x7, x9, x10, x11, x13 have negative association while x2, x4, x6, x8, x12 have positive association.

Table 2 Correlation matrix

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
y medv	1.00000	-0.40175 <.0001	0.39837 <.0001	-0.55153 <.0001	0.12126 0.0223	-0.47753 <.0001	0.73652 <.0001	-0.39869 <.0001	0.29423 <.0001	-0.45853 <.0001	-0.55432 <.0001	-0.54856 <.0001	0.34922 <.0001	-0.74901 <.0001
x1 crim	-0.40175 <.0001	1.00000	-0.18722 0.0004	0.37496 <.0001	-0.07156 0.1785	0.40489 <.0001	-0.16946 0.0014	0.32669 <.0001	-0.35289 <.0001	0.60395 <.0001	0.55540 <.0001	0.27328 <.0001	-0.34882 <.0001	0.42601 <.0001
x2 zn	0.39837 <.0001	-0.18722 0.0004	1.00000	-0.52665 <.0001	-0.01448 0.7856	-0.51123 <.0001	0.34416 <.0001	-0.55179 <.0001	0.65319 <.0001	-0.30078 <.0001	-0.31602 <.0001	-0.37300 <.0001	0.18132 0.0006	-0.43791 <.0001
x3 indus	-0.55153 <.0001	0.37496 <.0001	-0.52665 <.0001	1.00000	0.01200 0.8217	0.74492 <.0001	-0.44388 <.0001	0.62058 <.0001	-0.69288 <.0001	0.56081 <.0001	0.70216 <.0001	0.38502 <.0001	-0.35497 <.0001	0.64416 <.0001
x4 chas	0.12126 0.0223	-0.07156 0.1785	-0.01448 0.7856	0.01200 0.8217	1.00000	0.03563 0.5034	0.02949 0.5797	0.05425 0.3081	-0.05231 0.3257	-0.09062 0.0882	-0.11277 0.0337	-0.18661 0.0004	0.06180 0.2455	-0.01741 0.7437
x5 nox	-0.47753 <.0001	0.40489 <.0001	-0.51123 <.0001	0.74492 <.0001	0.03563 0.5034	1.00000	-0.35212 <.0001	0.72636 <.0001	-0.77177 <.0001	0.57237 <.0001	0.63751 <.0001	0.14199 0.0074	-0.38686 <.0001	0.64052 <.0001
x6 rm	0.73652 <.0001	-0.16946 0.0014	0.34416 <.0001	-0.44388 <.0001	0.02949 0.5797	-0.35212 <.0001	1.00000	-0.22367 <.0001	0.21869 <.0001	-0.22948 <.0001	-0.33017 <.0001	-0.36949 <.0001	0.11746 0.0269	-0.59240 <.0001
x7 age	-0.39869 <.0001	0.32669 <.0001	-0.55179 <.0001	0.62058 <.0001	0.05425 0.3081	0.72636 <.0001	-0.22367 <.0001	1.00000	-0.73168 <.0001	0.44095 <.0001	0.49269 <.0001	0.23268 <.0001	-0.27158 <.0001	0.61639 <.0001
x8 dis	0.29423 <.0001	-0.35289 <.0001	0.65319 <.0001	-0.69288 <.0001	-0.05231 0.3257	-0.77177 <.0001	0.21869 <.0001	-0.73168 <.0001	1.00000	-0.46847 <.0001	-0.52161 <.0001	-0.20271 0.0001	0.29597 <.0001	-0.53327 <.0001
x9 rad	-0.45853 <.0001	0.60395 <.0001	-0.30078 <.0001	0.56081 <.0001	-0.09062 0.0882	0.57237 <.0001	-0.22948 <.0001	0.44095 <.0001	-0.46847 <.0001	1.00000	0.89681 <.0001	0.45351 <.0001	-0.47495 <.0001	0.53031 <.0001
x10 tax	-0.55432 <.0001	0.55540 <.0001	-0.31602 <.0001	0.70216 <.0001	-0.11277 0.0337	0.63751 <.0001	-0.33017 <.0001	0.49269 <.0001	-0.52161 <.0001	0.89681 <.0001	1.00000	0.45695 <.0001	-0.46290 <.0001	0.59126 <.0001
x11 ptratio	-0.54856 <.0001	0.27328 <.0001	-0.37300 <.0001	0.38502 <.0001	-0.18661 0.0004	0.14199 0.0074	-0.36949 <.0001	0.23268 <.0001	-0.20271 0.0001	0.45351 <.0001	0.45695 <.0001	1.00000	-0.18319 0.0005	0.39197 <.0001
x12 b	0.34922 <.0001	-0.34882 <.0001	0.18132 0.0006	-0.35497 <.0001	0.06180 0.2455	-0.38686 <.0001	0.11746 0.0269	-0.27158 <.0001	0.29597 <.0001	-0.47495 <.0001	-0.46290 <.0001	-0.18319 0.0005	1.00000	-0.38077 <.0001
x13 lstat	-0.74901 <.0001	0.42601 <.0001	-0.43791 <.0001	0.64416 <.0001	-0.01741 0.7437	0.64052 <.0001	-0.59240 <.0001	0.61639 <.0001	-0.53327 <.0001	0.53031 <.0001	0.59126 <.0001	0.39197 <.0001	-0.38077 <.0001	1.00000

5 Model Building

First we build a standardized regression model to get an over view of the parameters. Make correlation transformation of the predictors as a form txi, i=1, 2….12, 13.From table 3, we can find that the tx3, tx7 and tx10 are not significant. T13, t6 and t8 are most important, and tx5, tx9, and tx11 are less important. T1, t2, t4, and t12 are lest important. This information can be used in added variable method.

Table 3 standard regression

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6.20437E-18	0.00145	0.00	1.0000	0
tx1	1	-0.11005	0.03587	-3.07	0.0023	1.72112
tx2	1	0.10736	0.04096	2.62	0.0092	2.24405
tx3	1	0.00495	0.05552	0.09	0.9289	4.12399
tx4	1	0.08815	0.02861	3.08	0.0022	1.09494
tx5	1	-0.24434	0.05640	-4.33	<.0001	4.25462
tx6	1	0.32802	0.04063	8.07	<.0001	2.20864
tx7	1	-0.00654	0.04944	-0.13	0.8949	3.26912
tx8	1	-0.33839	0.05328	-6.35	<.0001	3.79688
tx9	1	0.21694	0.07590	2.86	0.0045	7.70647
tx10	1	-0.12608	0.08623	-1.46	0.1446	9.94627
tx11	1	-0.21344	0.03713	-5.75	<.0001	1.84464
tx12	1	0.08767	0.03180	2.76	0.0061	1.35237
tx13	1	-0.37083	0.04867	-7.62	<.0001	3.16835

We will use several criteria, such as rsqure, adjrsq, cp, AIC and SBC, to help us select the best model. Stepwise, forward and backward selection also help us validate our choice.

Table 4 shows us the 10 models based on the criteria selection. I select the high light one with 10 variables. Because it has the smallest number of variables in the model with Adjust R-square increase in the corner, and CP closed to 10, and AIC and SBC are small enough.

Stepwise, backward and forward draw the same model with 10 predictors, with cx3, cx7 and cx10 eliminated. These selections method agree with my choice based on criterion.

Although criterion suggest us to choose cx1, cx2, cx4, cx5, cx6, cx8, cx9, cx11, cx12 and cx13, we should use added variable plots to determine if the variables should be added or should be added in a linear way or in a curved way. We begin with cx13 first into the model and add each variable each time into the model.

Table 3 Auto selection based on several criterions.

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
11	0.7369	0.7451	10.0256	1154.6859	1201.15127	cx1 cx2 cx4 cx5 cx6 cx8 cx9 cx10 cx11 cx12 cx13
12	0.7361	0.7451	12.0080	1156.6675	1207.00503	cx1 cx2 cx4 cx5 cx6 cx7 cx8 cx9 cx10 cx11 cx12 cx13
12	0.7361	0.7451	12.0175	1156.6774	1207.01494	cx1 cx2 cx3 cx4 cx5 cx6 cx8 cx9 cx10 cx11 cx12 cx13
10	0.7355	0.7430	10.7694	1155.5307	1198.12396	cx1 cx2 cx4 cx5 cx6 cx8 cx9 cx11 cx12 cx13
13	0.7354	0.7451	14.0000	1158.6592	1212.86886	cx1 cx2 cx3 cx4 cx5 cx6 cx7 cx8 cx9 cx10 cx11 cx12 cx13
11	0.7352	0.7435	12.1855	1156.9272	1203.39257	cx1 cx2 cx3 cx4 cx5 cx6 cx8 cx9 cx11 cx12 cx13
11	0.7348	0.7431	12.7123	1157.4717	1203.93713	cx1 cx2 cx4 cx5 cx6 cx7 cx8 cx9 cx11 cx12 cx13
12	0.7345	0.7435	14.1378	1158.8779	1209.21540	cx1 cx2 cx3 cx4 cx5 cx6 cx7 cx8 cx9 cx11 cx12 cx13
10	0.7320	0.7396	15.3823	1160.2626	1202.85587	cx1 cx4 cx5 cx6 cx8 cx9 cx10 cx11 cx12 cx13
10	0.7318	0.7394	15.6335	1160.5184	1203.11172	cx1 cx2 cx4 cx5 cx6 cx8 cx9 cx10 cx11 cx13

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	cx13			1	0.5518	0.5518	248.614	434.52	<.0001
2	cx6			2	0.0982	0.6499	119.282	98.72	<.0001
3	cx11			3	0.0319	0.6818	78.6392	35.16	<.0001
4	cx8			4	0.0126	0.6944	63.7958	14.42	0.0002
5	cx5			5	0.0206	0.7151	38.1735	25.29	<.0001
6	cx4			6	0.0085	0.7236	28.7805	10.72	0.0012
7	cx12			7	0.0054	0.7290	23.4978	6.97	0.0087
8	cx2			8	0.0049	0.7339	18.9715	6.34	0.0122
9	cx1			9	0.0031	0.7370	16.7832	4.11	0.0435
10	cx9			10	0.0060	0.7430	10.7694	8.02	0.0049
11	cx10			11	0.0021	0.7451	10.0256	2.76	0.0976
12		cx10		10	0.0021	0.7430	10.7694	2.76	0.0976

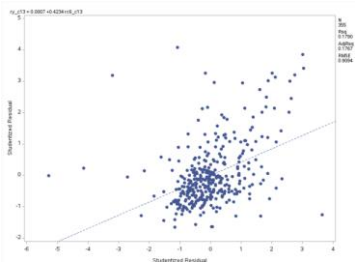
Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	cx3		12	0.0000	0.7451	12.0080	0.01	0.9289
2	cx7		11	0.0000	0.7451	10.0256	0.02	0.8943
3	cx10		10	0.0021	0.7430	10.7694	2.76	0.0976

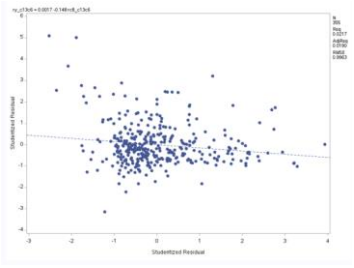
Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	cx13		1	0.5518	0.5518	248.614	434.52	<.0001
2	cx6		2	0.0982	0.6499	119.282	98.72	<.0001
3	cx11		3	0.0319	0.6818	78.6392	35.16	<.0001
4	cx8		4	0.0126	0.6944	63.7958	14.42	0.0002
5	cx5		5	0.0206	0.7151	38.1735	25.29	<.0001
6	cx4		6	0.0085	0.7236	28.7805	10.72	0.0012
7	cx12		7	0.0054	0.7290	23.4978	6.97	0.0087
8	cx2		8	0.0049	0.7339	18.9715	6.34	0.0122
9	cx1		9	0.0031	0.7370	16.7832	4.11	0.0435
10	cx9		10	0.0060	0.7430	10.7694	8.02	0.0049

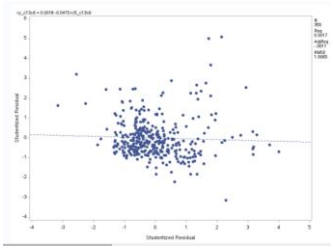
Figure 1 added variable into the model.



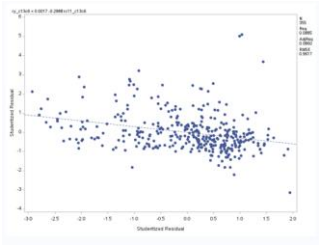
Cx6



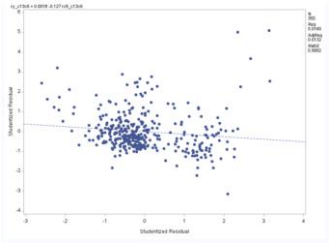
cx8



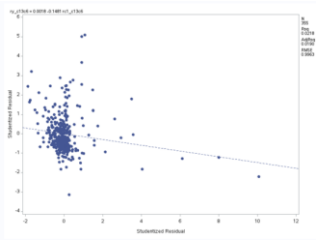
Cx5



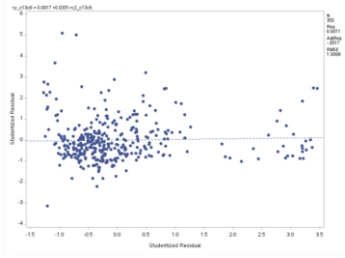
cx11



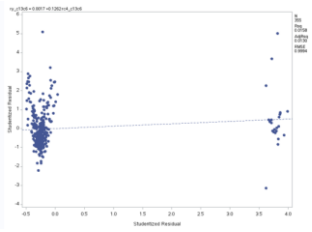
Cx9



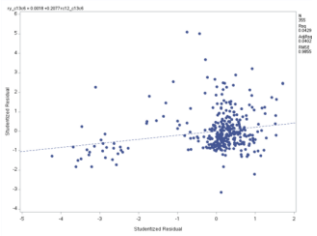
cx1



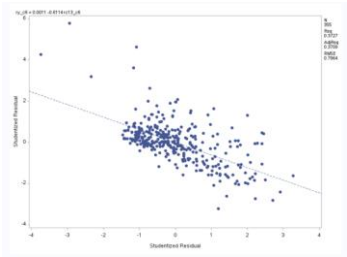
Cx2



cx4



Cx12



cx13 given cx6 in the model

From Figure 1 we can find that cx6, cx12 and cx13 should be stay in

the model. Cx8, cx5, cx11, cx9, cx2, and cx4 are negligible small effect on the model when cx6, cx12, and cx13 have already in the model, because the slop is so flat. The slop of cx1 looks deep. But when look into the plot, we find out that that slop is due to the points far away which maybe outlier. Neglect those point, the slop of cx1 is still flat. We also find that all the independent variables should be added in the model with approximate linear function. Further function form needs residual analysis.

Table 5 show the regression output from SAS. VIF ‘s show multicollinearity is not a problem in the model. The F value of the model is 205.61, and the p value is less than 0.001. This shows that the model is significant. Cx13 is negative associate with the response variable but cx6 and cx12 are positive associate with the response variable. All the relations with the response variable agree with the correlation matrix in table 2.

Take interaction term into the model. From table 6 we find out that R square increases to 0.724 from 0.637. We should take interaction into the model.

Table 5 preliminary regression output without interaction terms

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	18748	6249.35521	205.61	<.0001
Error	351	10668	30.39351		
Corrected Total	354	29416			

Root MSE	5.51303	R-Square	0.6373
Dependent Mean	22.39380	Adj R-Sq	0.6342
Coeff Var	24.61855		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	22.39380	0.29260	76.53	<.0001	0
cx13		1	-0.52878	0.05596	-9.45	<.0001	1.98628
cx6		1	5.37091	0.56408	9.52	<.0001	1.71036
cx12		1	0.01314	0.00350	3.76	0.0002	1.21373

Table 6 Preliminary SAS output with interaction terms

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	21302	3550.32794	152.27	<.0001
Error	348	8114.21871	23.31672		
Corrected Total	354	29416			

Root MSE	4.82874	R-Square	0.7242
Dependent Mean	22.39380	Adj R-Sq	0.7194
Coeff Var	21.56284		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	20.91240	0.31090	67.26	<.0001	0
cx13		1	-0.78523	0.05951	-13.20	<.0001	2.92804
cx6		1	3.56971	0.54142	6.59	<.0001	2.05395
cx12		1	0.00980	0.00457	2.14	0.0328	2.70112
cx13_cx6		1	-0.43326	0.04930	-8.79	<.0001	1.71311
cx13_cx12		1	-0.00042324	0.00046520	-0.91	0.3636	2.69489
cx6_cx12		1	-0.00024922	0.00616	-0.04	0.9677	1.76336

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	21281	5320.17069	228.88	<.0001
Error	350	8135.50359	23.24430		
Corrected Total	354	29416			

Root MSE	4.82123	R-Square	0.7234
Dependent Mean	22.39380	Adj R-Sq	0.7203
Coeff Var	21.52932		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	21.00373	0.28846	72.81	<.0001	0
cx13		1	-0.78561	0.05477	-14.34	<.0001	2.48841
cx6		1	3.62460	0.52089	6.96	<.0001	1.90709
cx12		1	0.00869	0.00312	2.14	0.0328	1.26332
cx13_cx6		1	-0.43948	0.04210	-10.44	<.0001	1.25346

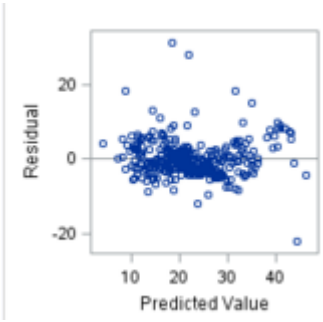
6 Residual Diagnostics

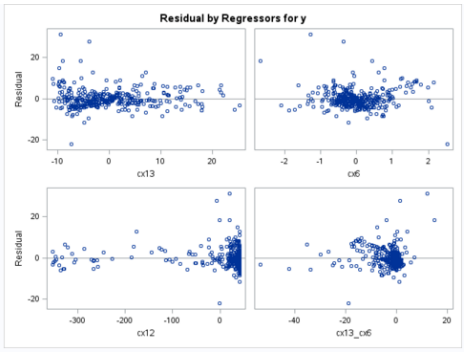
We have already build a final model with 3 predictor and one interaction term. Before analysis, we have to check the assumption of the model. The assumptions are

1. $E\{ \varepsilon_i \}=0$
2. ε_i Independent of ε_j
3. $\varepsilon_i \sim N(0,\sigma 2)$
4. $Var(\varepsilon_i) = \sigma 2$

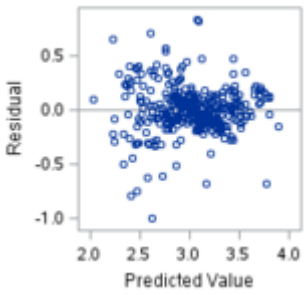
Figure 2

Before transformation





log transformation



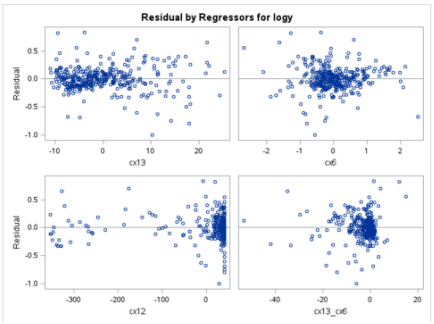
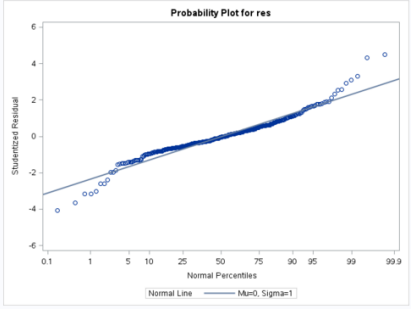
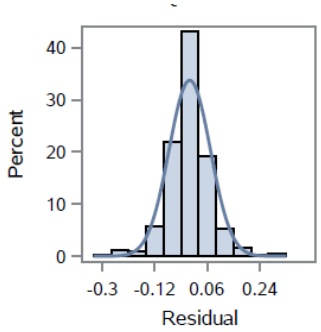


Figure3 normality plot and histogram





We can check residuals vs predicted value and predictors to check for independent and constant variance. From figure 2, on the left side we find out

that the residual favor on the positive side. On the right side, using box-cox transformation, lambda=0.19. After transformation and get the residual plot, right side from figure 2. The residuals disperse more averagely and unstructured. And also the plots of residuals vs predictors improve, unstructured and disperse averagely. This satisfies constant variance and independent on predictors and dependent variable, which means error terms are independent (It is not a time series problem). Look at the normality plot in figure 3. We can find that 95% points fall in the strait line, which proves to be approximate normal distribution. The histogram shows normal. Heavy two tails cannot be canceled, this may be because the data is exactly t distribution. Spline transformation can good fit the data with R square up to 85%. But it is out of the scope of linear regression.

Box-cox transformation on dependent variable also improves the R square by 1 percent and lowers MSE very much from 23.24 to 0.00509.

7 Outliers

7.1 Checking outliers

Using the criteria of hii> 2p/n to detect x outliers, where hii is the diagonal elements of the hat matrix, and where 2p/n=2*14/355=0.07887.

Using studentized deleted residuals to detect y outliers.

$$\text{The studentized deleted residual is } t_i = \frac{d_i}{\sqrt{s^2\{d_i\}}} = \frac{e_i}{(1-h_{i,i})}\sqrt{\frac{(1-h_{i,i})}{MSE_{(i)}}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{i,i})}}.$$

The appropriate Bonferroni critical value is t(1- α /2n;n-p-1). In our model, t(1-0.05/(2*n),n-p-1)=t(1-0.05/(2*355),355-14-1)=t(0.9999296,340)=-3.851. Take its absolute value is 3.851. the absolute value of data statistic greater than 3.851 should be considered to be y outliers.

Table 7

X outliers detected by hii.

Obs	id	cx13	cx6	cx12	dff	cd	hii
261	366	-5.586	-2.68905	-0.640	1.06136	0.21976	0.10298
267	375	25.264	-2.11205	41.560	0.17589	0.00620	0.19153
271	385	17.924	-1.88205	-69.510	-0.46760	0.04358	0.08901
290	413	21.664	-1.62205	-326.550	1.09325	0.23390	0.12082
292	415	24.274	-1.73105	-267.070	-0.56879	0.06452	0.14011

Table 8

Y outliers detected by studentized deleted residual

Obs	cx13	cx13_cx6	cx13_cx12	cx6_cx12	boxy	p	res	cd	hii	stu_dl	dff
262	-9.446	12.0913	-190.619	-25.8312	2.10283	1.78888	4.50355	0.19401	0.045645	4.63337	1.01330
265	-3.826	1.4349	28.542	2.7979	2.10283	1.79560	4.32261	0.03000	0.007964	4.43648	0.39751
284	10.274	-5.8258	304.418	-16.8015	1.35770	1.64577	-4.05956	0.03706	0.011120	-4.15270	-0.44036

Table 9
X and Y outliers detected by DFFITS

Obs	id	cx13	cx6	cx12	dff	cd	hii
260	365	-7.416	2.52995	-0.790	-0.92048	0.16351	0.05816
261	366	-5.586	-2.68905	-0.640	1.06136	0.21976	0.10298
262	369	-9.446	-1.28005	20.180	1.01330	0.19401	0.04564
265	373	-3.826	-0.37505	-7.460	0.39751	0.03000	0.00796
269	381	4.504	0.71795	41.560	-0.41195	0.03349	0.02875
271	385	17.924	-1.88205	-69.510	-0.46760	0.04358	0.08901
279	399	17.884	-0.79705	41.560	-0.54680	0.05825	0.02822
281	401	14.064	-0.26305	41.560	-0.50344	0.04952	0.02653
284	406	10.274	-0.56705	29.630	-0.44036	0.03706	0.01112
287	410	7.074	0.60195	-175.980	0.59847	0.06961	0.03110
290	413	21.664	-1.62205	-326.550	1.09325	0.23390	0.12082
292	415	24.274	-1.73105	-267.070	-0.56879	0.06452	0.14011

DFFITS is a diagnostic meant to show how influential a point is. The formula is

$$\mathrm{DFFITS} = \frac{\widehat{y_i} - \widehat{y_{i(i)}}}{s_{(i)}\sqrt{h_{ii}}}$$

The letters DF stand for the difference between the fitted value $\widehat{y_i}$ including ith case and the fitted value $\widehat{y_{i(i)}}$ excluding ith case. S(i) is the stander error of ith case, hii is the diagonal elements of hat matrix. Values larger than 2*sqrt(p/n) in absolute value are considered highly influential.

7.2 weighted Least squares

Instead of minimize the regular sum of square of residual. We can minimize the weighted least squares.

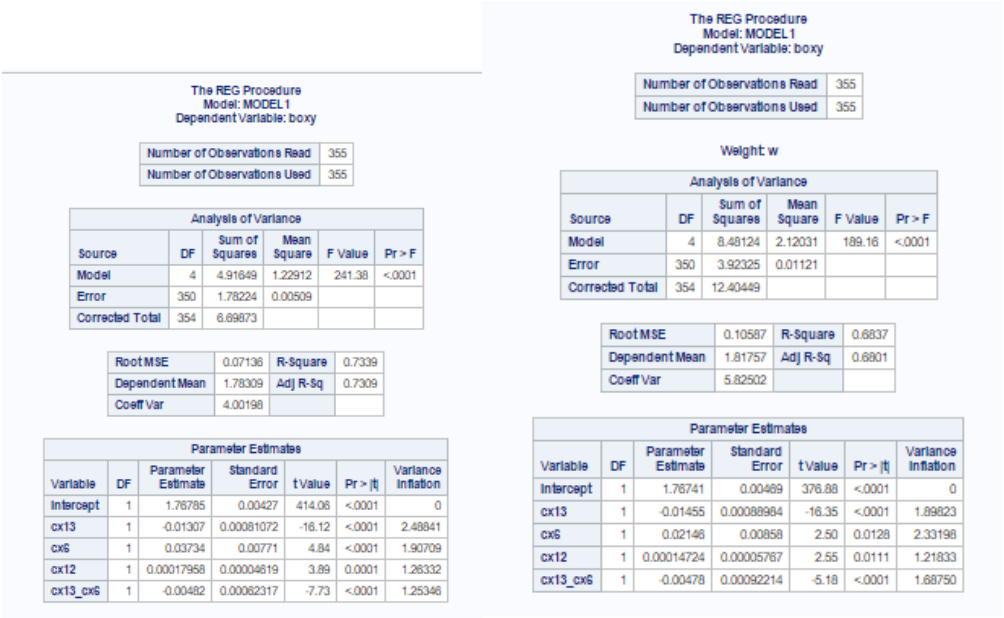
$$\mathrm{WSSE} = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} w_i \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_k x_{i,k} \right) \right]^2 \qquad w_i \propto \frac{1}{\sigma_i^2},$$

When σ_i^2 is unknown, we use s_i^2 to estimate. But the disadvantage of weighted least squares is obvious. If the estimate is biased or by small number of replicates, it would lead to bad model.

In our data set, see table 10, we find out that using weighted method makes the model worse. We should not admit weighted least squares

method.

Table 10
Unweighted and weighted method.



The REG Procedure
Model: MODEL1
Dependent Variable: boxy

Number of Observations Read	355
Number of Observations Used	355

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.91649	1.22912	241.38	<.0001
Error	350	1.78224	0.00509		
Corrected Total	354	6.69873			

Root MSE	0.07136	R-Square	0.7339
Dependent Mean	1.78309	Adj R-Sq	0.7309
Coeff Var	4.00198		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.76785	0.00427	414.06	<.0001	0
cx13	1	-0.01307	0.00081072	-16.12	<.0001	2.48841
cx6	1	0.03734	0.00771	4.84	<.0001	1.90709
cx12	1	0.00017958	0.00004619	3.89	0.0001	1.26332
cx13_cx6	1	-0.00482	0.00062317	-7.73	<.0001	1.25346

The REG Procedure
Model: MODEL1
Dependent Variable: boxy

Number of Observations Read	355
Number of Observations Used	355

Weight: w

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	8.48124	2.12031	189.16	<.0001
Error	350	3.92325	0.01121		
Corrected Total	354	12.40449			

Root MSE	0.10587	R-Square	0.6837
Dependent Mean	1.81757	Adj R-Sq	0.6801
Coeff Var	5.82502		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.76741	0.00469	376.88	<.0001	0
cx13	1	-0.01455	0.00088984	-16.35	<.0001	1.89823
cx6	1	0.02146	0.00858	2.50	0.0128	2.33198
cx12	1	0.00014724	0.00005767	2.55	0.0111	1.21833
cx13_cx6	1	-0.00478	0.00092214	-5.18	<.0001	1.68750

7.3 comparing delete outliers and keep outliers

Since we cannot remedy the effect or outlier using weighted least squares method. We cannot simply delete these outliers. Compare the difference of the model parameters between deleting and keeping the outliers. In table 11, we find out that deleting only y outlier improve R square, but there are no other difference between deleting only Y outliers and deleting both Y and X outliers. Comparing table 11 and table 10, we find out that the parameter of cx6 increase slightly from 0.037 to 0.045 and R square increases only 3%. So we conclude that the outliers do not impact the model too much.

Table 11

Delete y outliers

The REG Procedure
Model: MODEL1
Dependent Variable: boxy

Number of Observations Read	352
Number of Observations Used	352

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.81791	1.20448	279.52	<.0001
Error	347	1.49527	0.00431		
Corrected Total	351	6.31318			

Root MSE	0.06564	R-Square	0.7632
Dependent Mean	1.78249	Adj R-Sq	0.7604
Coeff Var	3.68272		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.76649	0.00395	447.49	<.0001	0
cx13	1	-0.01211	0.00075624	-16.02	<.0001	2.53089
cx6	1	0.04583	0.00721	6.36	<.0001	1.94435
cx12	1	0.00020267	0.00004259	4.76	<.0001	1.26898
cx13_cx6	1	-0.00493	0.00057389	-8.59	<.0001	1.23618

delete x and y outliers

Model: MODEL1
Dependent Variable: boxy

Number of Observations Read	350
Number of Observations Used	350

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.80520	1.20130	247.37	<.0001
Error	345	1.67541	0.00486		
Corrected Total	349	6.48061			

Root MSE	0.06969	R-Square	0.7415
Dependent Mean	1.78510	Adj R-Sq	0.7385
Coeff Var	3.90380		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.76725	0.00432	408.66	<.0001	0
cx13	1	-0.01249	0.00082532	-15.13	<.0001	2.41391
cx6	1	0.04522	0.00852	5.31	<.0001	2.12580
cx12	1	0.00020235	0.00004742	4.27	<.0001	1.31043
cx13_cx6	1	-0.00475	0.00080449	-5.90	<.0001	1.34178

8 Final model

We reach the final model that include cx13, cx6, cx12 and the interaction between cx13 and cx6. The parameter of c12 is negligible small. Delete c12 in our model R square does not decrease much. And other parameters do not change much. So for interpretation better we can delete c12 and reach our final model only include cx13, cx6 and the interaction between them.

The ANOVA table show that the model is significant under the P-value of F test less than 0.0001. The P-value of each parameter is less than or equal to 0.0001. R square is 0.7225, which is good fit of the model. 72.25 of the variation of response variable is explained by the model. We also get the 95% confidence limits of the parameters. Notice that here we have center the predictor, so the parameters themselves explain the average effect of increasing one unit in the center variable has the effect on the box-cox transformation of the response variable.

Table 12 final model

Include cx12

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.91649	1.22912	241.38	<.0001
Error	350	1.78224	0.00509		
Corrected Total	354	6.69873			

Root MSE	0.07136	R-Square	0.7339
Dependent Mean	1.78309	Adj R-Sq	0.7309
Coeff Var	4.00198		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	1.76785	0.00427	414.06	<.0001	1.75945	1.77625
cx13	1	-0.01307	0.00081072	-16.12	<.0001	-0.01467	-0.01148
cx6	1	0.03734	0.00771	4.84	<.0001	0.02218	0.05251
cx12	1	0.00017958	0.00004619	3.89	0.0001	0.00008874	0.00027041
cx13_cx6	1	-0.00482	0.00062317	-7.73	<.0001	-0.00604	-0.00359

delete cx12

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4.83950	1.61317	304.55	<.0001
Error	351	1.85922	0.00530		
Corrected Total	354	6.69873			

Root MSE	0.07278	R-Square	0.7225
Dependent Mean	1.78309	Adj R-Sq	0.7201
Coeff Var	4.08167		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	1.76633	0.00434	407.33	<.0001	1.75780	1.77486
cx13	1	-0.01446	0.00074258	-19.47	<.0001	-0.01592	-0.01300
cx6	1	0.03103	0.00769	4.04	<.0001	0.01591	0.04615
cx13_cx6	1	-0.00530	0.00062298	-8.51	<.0001	-0.00652	-0.00407

9 Validation of the model

We use the test data split from the raw data to get the predictive power of our model. We use the exact method of our model building the get the ANOVA table from the test data and compare the difference. We find out that the parameter is similar except cx12. Cx12 becomes insignificant and only a half of the parameter of c12 in our model. This reinforces our consideration of deleting c12 in our final model. These evidences prove that our model has high predictive power.

We also can obtain MSPR(mean square prediction error) to check our model. If MSPR is very closed to the MSE from our model, then our model has good predictive power.

Table 13 validation model

Include cx12

Number of Observations Used	151

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.18850	0.54712	128.66	<.0001
Error	146	0.62084	0.00425		
Corrected Total	150	2.80934			

Root MSE	0.06521	R-Square	0.7790
Dependent Mean	1.79023	Adj R-Sq	0.7730
Coeff Var	3.64256		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.77371	0.00603	294.12	<.0001
cx13	1	-0.01495	0.00109	-13.70	<.0001
cx6	1	0.03677	0.00908	4.05	<.0001
cx12	1	0.00008838	0.00006405	1.38	0.1697
cx13_cx6	1	-0.00580	0.00101	-5.76	<.0001

final model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.18040	0.72680	169.87	<.0001
Error	147	0.62894	0.00428		
Corrected Total	150	2.80934			

Root MSE	0.06541	R-Square	0.7761
Dependent Mean	1.79023	Adj R-Sq	0.7716
Coeff Var	3.65375		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.77289	0.00602	294.52	<.0001
cx13	1	-0.01544	0.00103	-14.93	<.0001
cx6	1	0.03512	0.00903	3.89	0.0002
cx13_cx6	1	-0.00608	0.00098651	-6.17	<.0001

MSPR= $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2)/n^*$

Y_i is the value of the response variable in the ith validation case.

$Y(hat)_i$ is the predicted value of the ith validation case based on the model building data set.

n* is the number of cases in the validation set.

In our test data set, we get MSPR=0.00447, which is very closed the MSE=0.0053 from our final model. This indicates that our model has high predictive power.

Table 14
MSPR(only shows 15 observation)

Obs	id	p	model_p	mspr
1	6	1.89098	1.86815	.004470548
2	10	1.69968	1.70256	.004470548
3	13	1.71626	1.71577	.004470548
4	28	1.69756	1.70114	.004470548
5	29	1.77302	1.76895	.004470548
6	30	1.79321	1.78734	.004470548
7	32	1.75559	1.75091	.004470548
8	42	1.92468	1.89840	.004470548
9	43	1.86165	1.84193	.004470548
10	48	1.67710	1.68349	.004470548
11	51	1.74678	1.74255	.004470548
12	58	1.94462	1.91539	.004470548
13	59	1.84503	1.82765	.004470548
14	62	1.73399	1.73162	.004470548
15	63	1.86875	1.84952	.004470548

10 Conclusion

The purpose of this analysis is to determine what important factors can predict the median value of housing in Boston. In our final model, we find out that LSTAT (Proportion of lower status of the population), RM (average number of rooms per dwelling) and their interaction contain much information of the median value of house. Decrease the proportion of lower status of the population and increase the average number of rooms can increase the median value of the house in Boston. Although other predictors can predict the median value of house, but they contribute little when comparing to LSTAT and RM. Weighted least squares has its own disadvantage which assume the variance is known or easy to assess. Biased estimate would lead to bad model. Our paper only concerns the multiple linear regression method, although other methods can get a better prediction, such as

spline transformation of predictors and adding higher order of predictors.

Appendix 1

Kutner, Michael H. Applied linear statistical models. Boston: McGraw-Hill Irwin, 2005.
Sample project, author unknown
Data set is taken from UCI website
https://archive.ics.uci.edu/ml/datasets/Housing

Appendix 2
SAS code

```
data rawdat;
infile  "F:\course\510\hw\project\housing.data";
input id x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 y;
     label x1="crim"
          x2="zn"
          x3="indus"
          x4="chas"
          x5="nox"
          x6="rm"
          x7="age"
          x8="dis"
          x9="rad"
          x10="tax"
          x11="ptratio"
          x12="b"
          x13="lstat"
          y="medv";
          datalines;
run;
proc surveyselect    data=rawdat out=split    rate=0.7 outall;
```

```
run;
proc sql;
create table build as select * from split where selected=1;
quit;
proc sql;
create table test as select * from split where selected=0;
quit;
proc print data=build;
var x1-x13 y;
run;
proc print data=test;
var x1-x13 y;
run;

proc corr data=build;
var y x1-x13;
run;
proc sql;
create table a as
select *,       (y-mean(y))/(std(y)*sqrt(count(y)-1)) as ty,
                    (x1-mean(x1))/(std(x1)*sqrt(count(x1)-1)) as tx1,
                    (x2-mean(x2))/(std(x2)*sqrt(count(x2)-1)) as tx2,
                    (x3-mean(x3))/(std(x3)*sqrt(count(x3)-1)) as tx3,
                    (x4-mean(x4))/(std(x4)*sqrt(count(x4)-1)) as tx4,
                    (x5-mean(x5))/(std(x5)*sqrt(count(x5)-1)) as tx5,
                    (x6-mean(x6))/(std(x6)*sqrt(count(x6)-1)) as tx6,
                    (x7-mean(x7))/(std(x7)*sqrt(count(x7)-1)) as tx7,
                    (x8-mean(x8))/(std(x8)*sqrt(count(x8)-1)) as tx8,
                    (x9-mean(x9))/(std(x9)*sqrt(count(x9)-1)) as tx9,
                    (x10-mean(x10))/(std(x10)*sqrt(count(x10)-1)) as tx10,
                    (x11-mean(x11))/(std(x11)*sqrt(count(x11)-1)) as tx11,
                    (x12-mean(x12))/(std(x12)*sqrt(count(x12)-1)) as tx12,
                    (x13-mean(x13))/(std(x13)*sqrt(count(x13)-1)) as tx13,
                    (x1-mean(x1)) as cx1,
                    (x2-mean(x2)) as cx2,
                    (x3-mean(x3)) as cx3,
                    (x4-mean(x4)) as cx4,
                    (x5-mean(x5)) as cx5,
                    (x6-mean(x6)) as cx6,
                    (x7-mean(x7)) as cx7,
                    (x8-mean(x8)) as cx8,
                    (x9-mean(x9)) as cx9,
                    (x10-mean(x10)) as cx10,
                    (x11-mean(x11)) as cx11,
```

```
                    (x12-mean(x12)) as cx12,
                    (x13-mean(x13)) as cx13
from build;
proc reg data=a;
model ty=tx1-tx13/vif;
run;
proc reg data=a;
model y=cx1-cx13/vif selection= adjrsq aic sbc cp rsquare best=10;
run;
proc reg data=a;
model y=cx1-cx13/vif selection=stepwise sle=0.25 sls=0.05;
run;
proc reg data=a;
model y=cx1-cx13/vif selection=backward sls=0.05;
run;
proc reg data=a;
model y=cx1-cx13/vif selection=forward sle=0.05 sls=0.05;
run;

/*using added method to check. start from c13*/
proc reg data=a noprint;
model y cx6=cx13 ;
output out=res student=ry_c13 rc6_c13;
run;
proc reg data=res noprint;
model ry_c13=rc6_c13;
symbol1 i=sm70s v=dot;
plot ry_c13*rc6_c13;
run;
/*cx6 is important givin cx13 in the model*/
/*check cx13 given cx6 in the model*/
proc reg data=a noprint;
model y cx13= cx6;
output out=res student=ry_c6 rc13_c6;
run;
proc reg data=res;
model ry_c6=rc13_c6;
symbol1 i=sm70s v=dot;
plot ry_c6*rc13_c6;
run;

/*cx6 is important givin cx13 is in the model, add cx8*/
proc reg data=a noprint;
model y cx8=cx13 cx6;
```

```
output out=res student=ry_c13c6 rc8_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc8_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc8_c13c6;
run;
/*cx8 is not imporatn givin cx13 and cx6 are in the model*/
/*add cx5*/
proc reg data=a noprint;
model y cx5=cx13 cx6;
output out=res student=ry_c13c6 rc5_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc5_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc5_c13c6;
run;
/*cx5 is not important*/
/*add cx11*/

proc reg data=a noprint;
model y cx11=cx13 cx6;
output out=res student=ry_c13c6 rc11_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc11_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc11_c13c6;
run;

/*cx11 is not important*/
/*add cx9*/
proc reg data=a noprint;
model y cx9=cx13 cx6;
output out=res student=ry_c13c6 rc9_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc9_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc9_c13c6;
run;
/*cx9 is not important*/
/*add cx1*/
```

```
proc reg data=a noprint;
model y cx1=cx13 cx6;
output out=res student=ry_c13c6 rc1_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc1_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc1_c13c6;
run;
/*cx1 is not important*/
/*add cx2*/
proc reg data=a noprint;
model y cx2=cx13 cx6;
output out=res student=ry_c13c6 rc2_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc2_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc2_c13c6;
run;
/*cx2 is not important*/
/*add cx4*/
proc reg data=a noprint;
model y cx4=cx13 cx6;
output out=res student=ry_c13c6 rc4_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc4_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc4_c13c6;
run;
/*cx4 is not important*/
/*add cx12*/
proc reg data=a noprint;
model y cx12=cx13 cx6;
output out=res student=ry_c13c6 rc12_c13c6;
run;
proc reg data=res noprint;
model ry_c13c6=rc12_c13c6;
symbol1 i=sm70s v=dot;
plot ry_c13c6*rc12_c13c6;
run;
```

```
proc reg data=a;
model y=cx13 cx6 cx12/vif;
run;
data ac;
set a;
cx13_cx6=cx13*cx6;
cx13_cx12=cx13*cx12;
cx6_cx12=cx6*cx12;
run;
proc reg data=ac;
model y=cx13 cx6 cx12 cx13_cx6 cx13_cx12 cx6_cx12/vif;
run;
proc reg data=ac;
model y=cx13 cx6 cx12 cx13_cx6/vif;
output out=res r=res p=p;
run;
proc gplot data=res;
plot res*p;
plot res*(cx13 cx6 cx12 cx13_cx6);
run;
proc transreg data=ac;
model boxcox(y/lambda=-3 to 3 by 0.01)=identity(cx13 cx6 cx12);
run;

data ac;
set ac;
boxy=y**0.19;
run;
proc reg data=ac;
model boxy=cx13 cx6 cx12 cx13_cx6/vif influence;
output out=res1 student=res p=p rstudent=stu_dl h=hii cookd=cd dffits=dff ;
run;

proc univariate data=res1;
var res;
probplot/normal(mu=0 sigma=1 color=red);
run;
```

```
/*exam outliner*/
/*2p/n=2*14/355=0.07887*/
proc print data=res1;
var id cx13 cx6 cx12 dff cd hii;
where hii>0.07887;
run;
/*t(1-0.05/(2*n),n-p-1)=t(1-0.05/(2*355),355-14-1)=t(0.9999296,340)=-3.851*/
proc print data=res1;
where stu_dl>3.851 or stu_dl<-3.851;
run;

/*2*sqrt(p/n)=2*sqrt(14/355)=0.3972*/
proc print data=res1;
var id cx13 cx6 cx12 dff cd hii;
where dff>0.3972 or dff<-0.3972;
run;

/*F(0.05,14,355)=0.47*/
proc print data=res2;
var id cx13 cx6 cx12 dff cd hii;
where cd>0.47;
run;
/*weighted method*/
data wls;
set res1;
absr=abs(res);
run;
proc reg data=wls;
model absr=cx13 cx6 cx12 cx13_cx6;
output out=weigh p=s;
run;
data weighted;
set weigh;
w=1/(s**2);
run;
proc reg data=weighted;
weight w;
model boxy=cx13 cx6 cx12 cx13_cx6/vif influence;
run;

proc reg data=weighted;
weight w;
model boxy=cx13 cx6 cx12 cx13_cx6;
run;
```

```
/*delete    y outlier*/
data res2;
set res1;
where -3.851<stu_dl<3.851;
run;
proc reg data=res2;
model boxy=cx13 cx6 cx12 cx13_cx6/vif;
output out=res3 student=res1 p=p1;
run;
proc gplot data=res3;
plot res1*p1;
run;

/*delet x y outlier*/
data res4;
set res1;
where -3.851<stu_dl<3.851;
where    -0.3972<dff<0.3972;
where hii<0.07887;
run;
proc reg data=res4;
model boxy=cx13 cx6 cx12 cx13_cx6/vif;
output out=res5 student=res1 p=p1;
run;
proc gplot data=res5;
plot res1*p1;
run;

/*final model*/
proc reg data=ac;
model boxy=cx13 cx6 cx12 cx13_cx6/p cli clm clb;
run;


/*model validation*/
proc sql;
create table tes as
select *,
                    (y**0.19) as boxy,
                    (x1-mean(x1)) as cx1,
                    (x2-mean(x2)) as cx2,
                    (x3-mean(x3)) as cx3,
                    (x4-mean(x4)) as cx4,
```

```
                (x5-mean(x5)) as cx5,
                (x6-mean(x6)) as cx6,
                (x7-mean(x7)) as cx7,
                (x8-mean(x8)) as cx8,
                (x9-mean(x9)) as cx9,
                (x10-mean(x10)) as cx10,
                (x11-mean(x11)) as cx11,
                (x12-mean(x12)) as cx12,
                (x13-mean(x13)) as cx13

from test;
quit;
data b;
set tes;
cx13_cx6=cx13*cx6;
run;
proc reg data=b;
model boxy=cx13 cx6 cx12 cx13_cx6;
run;
proc reg data=b;
model boxy=cx13 cx6 cx13_cx6;
output out=valid    p=p;
run;

PROC SQL;
CREATE TABLE MSPR AS
SELECT *,
      (1.76785-0.01307*cx13+0.03734*cx6 -0.00482*cx13_cx6) as model_p,
       (mean((boxy-model_p)**2)) as mspr
      from valid;
      quit;

proc print data=mspr;
var id p model_p mspr;
run;
```